Research on Intelligent English Oral Training System in Mobile Network

Fen Zhu
Foreign Languages School, Luoyang Institute of Science and Technology, Luoyang, Henan, 471000, China
E-mail: fenzhu_lit@163.com

Technical Paper

Keywords: mobile network, Android system, spoken English, resonance peak, evaluation

Received: March 13, 2018

With the rapid development of mobile networks, mobile learning, as a new learning form, is gradually accepted by people. Based on the Android mobile platform, this paper designed a spoken English training system that could be applied to mobile network equipment from the aspects of speech recognition, pronunciation scoring and function setting. Based on the characteristics of the Android system, this paper selected the MEL cepstrum coefficient as the feature parameters to speech recognition, and introduced the dynamic time neat algorithm as the matching algorithm of the speech recognition pattern to make speech recognition more suitable for mobile Internet devices. Besides, the voice formant was used as a reference for oral scores and the scoring method based on single reference template was adopted. Finally, the spoken English training system was developed under the eclipse integration environment. The test results showed that the success rate of voice input was over 98%, and the accuracy rate of spoken voices of monophthong words, diphthong words and polysyllabic words was 97.15%, 94.96% and 93.62% respectively, suggesting that the system could accurately input and score English learners' spoken English, and assist English pronunciation.

Povzetek: Prispevek se ukvarja z mobilnim učenjem angleščine na sistemih z Androidom..

1 Introduction

With the deepening of economic globalization, communication between China and other counties has become increasingly frequent. Therefore English which is the most extensively applied language worldwide has gradually been an indispensable tool in daily life and work, and moreover many English training institutions and learning tools have emerged. But the traditional learning mode, i.e. face-to-face teaching mode in training institutions, usually cannot achieve a good result in spoken English, which contributes to the large difference of pronunciation between English and Chinese. People who grow up in Chinese environment will make the mistake of pronunciation unconsciously when learning oral English. Moreover English teachers who have correct pronunciation and are able to guide pronunciation are lack of in China. Time and environment for spoken English practice are also not enough.

With the rapid development of mobile information technology, mobile network terminals such as smartphone and panel personal computer have almost covered every aspect of our life. Smartphone based oral English training software is more convenient and practical compared to the traditional teaching mode and can effectively avoid the shortcomings of the traditional teaching mode. Mobile network device based mobile learning has been extensively studied. Wang et al. [1] found that computer corpus based teaching mode was more effective than the traditional teaching mode. Alamer et al. [2] designed and develop mobile Web technology and API based lightweight language learning management system. The system aimed to allow language students to view and download learning content on their phones and complete interactive tasks designed by teachers. Milutinovic [3] et al. proposed a mobile adaptive language learning model, whose main goal was to improve the mobile language learning process using adaptive technology. The proposed model was designed to take advantage of unique opportunities to transfer learning content in real learning situations. Taking Android smartphone as the application platform, this study aimed to build an intelligent spoken English training system that could be used on mobile network devices.

2 Mobile learning

Mobile learning [5] refers to the use of portable mobile communication equipment and technology so that learners can choose their preferred way to study any time and place. Compared with the time fixed English classroom learning mode, mobile learning has extensiveness, timeliness and interactivity features, giving learners a more relaxed and pleasant learning experience. In addition, the multimedia combination of audio, text, video, image and animation makes mobile learning more vivid. Mobile language learning enables

learners to have more learning options, to make full use of fragmented time, and to be efficient and flexible.

3 Intelligent spoken English training system design

3.1 Speech recognition

3.1.1 Speech signal preprocessing

(1) Speech signal digitization

Speech signal can be analyzed and processed by computer through digital conversion. This paper uses the headset of Android phone as the input device of voice signal, and uses the Audio Record Wizard [6] of Android system to collect the underlying data. According to Nyquist frequency theorem, the sampling frequency of 7000 Hz is used to collect the speech signal.

(2) Pre-emphasis

In order to eliminate the influence of mouth and nose radiation, speech signals are usually pre-emphasized by a first-order high-pass filter [7]. Pre-emphasis refers to improving the resolution of the high-frequency part of speeches by emphasizing the high-frequency part of speeches based on the difference between signal properties and noise properties. Usually pre-emphasis is realized using first-order FIR high-pass digital filter [16]. The formula used by the filter is shown below.

$$H(x)=1-\varepsilon x^{-1}, \tag{1}$$

Where ε refers to the pre-emphasis coefficient and is set to 0.98 in this study. Set the speech signal at the nth time point to be $s(n)$, then the weighted signal is:

$$s_2(n)=s(n)-\varepsilon s(n-1), \tag{2}$$

Where $s_2(n)$ refers to the speech signal after pre-emphasis and $s(n-1)$ refers to the last filter output value.

(3) Windowing processing

In order to ensure continuous and complete voice signals in each frame, a window function is generally multiplied before processing each frame of speech [8]. This paper uses Hamming window function to window the signal, with the formula as follows:

$$w(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)], & 0\le n\le N-1\\ 0, & n=othervalue\end{cases}. \tag{3}$$

(4) Endpoint detection

According to the characteristics of the Android platform, this paper uses the combination of short time average energy [9] and short-time zero-crossing rate to detect the endpoint. The short-term average energy is calculated as follows: set the short time average energy and frame length of the n-th frame of speech signal $s_n(h)$ to E_n and N respectively, then the calculation formula is as follows:

$$E_n=\sum_{m=0}^{N-1}x_n^2(h),0\le h\le N-1 \tag{4}$$

According to the size of the short-term energy, the learner's voice and noise can be distinguished, and high energy signal is the speech signal. However, this method is less stable under low SNR conditions. Therefore, it is necessary to use short-time zero-rate method. Set the speech signal to be $x_n(m)$, then the short-time zero-crossing rate is:

$$z_n=\frac{1}{2}\sum_{m=0}^{N-1}|\mathrm{sgn}[x_n(h)]-\mathrm{sgn}[x_n(h-1)]\quad \mathrm{sgn}[x]=\begin{cases}1 & (x\ge 0)\\ -1 & (x\le 0)\end{cases}, \tag{5}$$

Where sgn[ ] refers to the sign function. According to the low frequency band of voiced sound energy and the high frequency band of voiceless sound energy, the zero-crossing rate of the speaker is stable relative to the ambient noise and the sound segment can be clearly identified.

3.1.2 Extraction of speech signal features

Feature extraction [17] was performed after the preprocessing to highlight the data features of pattern matching, improve recognition rate, compress information and reduce computation load and storage. The commonly used feature parameters include Mel-frequency cepstral coefficient (MFCC) which has strong recognition performance and anti-noise capacity, linear predictive coefficient which has small computer load but general efficacy and accent sensitivity parameter which has favorable performance in recognition the middle frequency band of signals.

In this system, Mel Frequency Cepstrum Coefficient (MFCC) [10] is used as the characteristic parameter of oral training. MEL scale and frequency have the following relationship:

$$f_{mel}=2595\ln(1+f/700), \tag{6}$$

Where f refers to the actual frequency of the signal.

Fourier transform [11] is performed on each frame of speech signal after preprocessing to obtain the signal spectrum. Then, the spectrum square is cut off, Mel band-pass filter is applied for filtering, all of the filter outputs undergo logarithm calculation, and then discrete cosine transform is made on DCT to obtain MFCC, the process is shown in Figure 1.

$$C(n)=\sqrt{\frac{2}{N}}\log w(l)\cos\left\{\left(l-\frac{1}{2}\right)\frac{n\pi}{L}\right\},(n=1,2,...,p)(l=1,2,...,L), \tag{7}$$

Where L refers to the number of filters, $w(l)$ refers to the output of each triangle filter, N refers to the length of each frame, and p refers to the order of parameters.

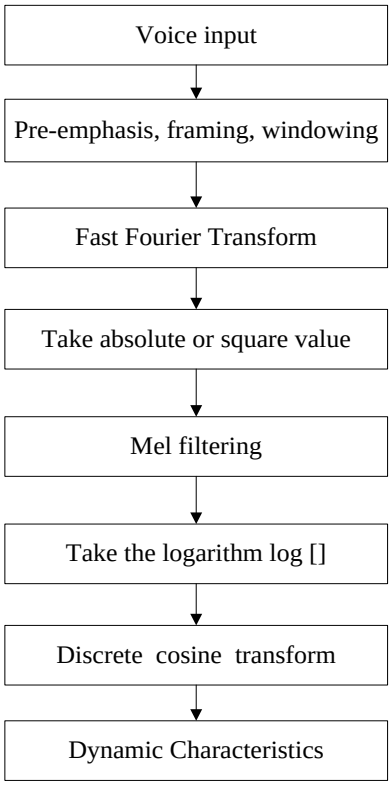

Figure 1: MFCC feature extraction process.

3.2 Speech signal pattern matching

In this paper, Dynamic Time Warping (DTW) [12] is used to match the characteristics of speech signals. Firstly, set the eigenvector sequence of the standard template to be $B=\{B(1),B(2),\ldots,B(m),\ldots,B(M)\}$, where M is the total speech frame number, m is the time series label of the signal frame and $B(m)$ is the eigenvector of the m-th frame.

The eigenvector sequence of the speech test template is $T=\{T(1),T(2),\ldots,T(n),\ldots,T(Q)\}$, where Q is the number of frames, n is the sequence number of the speech in the template and $T(n)$ is the eigenvector of the nth frame. The similarity between the test template and the standard template is represented by vector distance, and the similarity decreases with the increase of vector distance. Euclidean distance [13] is used to represent the distance between $T(n)$ and $B(m)$, as follows:

$$d[T(n),B(m)]=\sum_{i=1}^{p}(t_i-b_i)^2 \quad (8)$$

Where t_i refers to the eigenvector of the i-th dimension of $T(n)$, b_i refers to the eigenvector of the i-th dimension of $B(m)$. The dynamic time warping is to map the time axis n of the speech test template to the time axis m of the standard template to obtain the minimum vector distance of the template, as follows:

$$D=\min_{w(n)}\sum_{n=1}^{N}d[T(n),b(m)] \quad (9)$$

Dynamic time warping generally requires finding a path which goes through each intersection with the distance measure sum of the intersection at the path minimized. Generally, constraint conditions are given:

Boundary condition:

$$w(1)=1, w(N)=M \quad (10)$$

Continuity condition:

$$w(n+1)-w(n)=\begin{cases}0,1,2 & w(n)\neq w(n-1)\\ 1,2 & w(n)=w(n-1)\end{cases} \quad (11)$$

With the above two conditions met and the frame distance accumulated sum the minimum, the optimal path $m=w(n)$ is sought as follows: starting from (1, 1), backstepping is repeated until (N, M) to find the optimal matching path. $D(N,M)$ refers to template distance of the matching path and the minimum matching distance is $D\min(N,M)$, which is taken as the measuring criterion for the similarity matching degree between templates.

3.3 Pronunciation scoring

Firstly, the average matching distance of frames is calculated:

$$\bar{d}=\frac{D(N.M)}{N}, \quad (12)$$

Where $D(N,M)$ refers to the total matching distance of the test templates, N refers to the frame length of the test templates. When selecting the average frame matching distance, the effect of the speech length is eliminated. In terms of scoring, this paper proposes a scoring method based on the single reference template. The range of pronunciation score is 0~100, and the scoring method is as follows:

$$score=\frac{100}{1+e(d)^f}, \quad (13)$$

Where d refers to the average frame matching degree, and e and f are the scoring parameters obtained based on the experience of spoken English teachers and matching distance.

3.4 Scoring parameter selection

In this study, the formant was taken as a criterion to evaluate the learner's spoken language pronunciation, and the learners' spoken English pronunciation quality was judged by the similarity contrast between the pronunciation formant of the test model and the standard model. Formant refers to the areas where energies are concentrated in the speech spectrum and it reflects the physical characteristics of the resonant cavity. In the process of producing vowels and consonants in the oral cavity, the harmonic vibration frequency of the sound is regulated by the sound cavity, which is strengthened or attenuated irregularly, and the region with high degree of enhancement forms the resonance peak. In the spectrum

of vowels, the first three resonant peaks play a key role in the quality of sound. The first two resonant peaks are particularly sensitive to the height of the tongue position. The higher the first resonance peak, the lower the tongue position, and the second and third formants also have a certain relationship with the tongue position, but the relationship between them is not particularly prominent. Therefore, the first resonance peak is chosen as the judging basis for the pronunciation quality. In this paper, the resonance peak is extracted using linear prediction method [14]. Regarding the sound channel as a resonant cavity, then the resonant peak is the resonant frequency of the wall.

3.5 Function and interface design

The oral English training system based on the Android smartphone platform can provide effective feedback to learners' oral English pronunciation through animation, audio, video and image forms. The function design of the system is as follows: First of all, the system should have standard pronunciation audios and videos to guide learners, and introduce the key points of English pronunciation and tongue type in the form of pictures and texts. Before establishing the system, spoken phonetic materials such as phonetic symbols, words, and sentences need to be collected. Folders of pictures, videos and texts should be established separately for system access. We use AudioTrack for audio and video playback, specifically, class method for speech signal playback and Video View class method in Android SDK for video playback. Secondly, the system should be able to prompt the learner to read the words and phrases, record and play back the voice signals, create a cache folder, and record the recorded voice signals according to the MP3 format. AudioRecond class method is used to record voice signals, and the sampling frequency is set to 8000Hz, channel mono, 16-bit sampling bits.

Then, the system uses the speech recognition and related algorithms to score learners' spoken pronunciations and establish a spoken appraisal folder. The Shared Preferences component of the Android system is used to store the learner's spoken rating results. Finally, the system should have the function of comparing learner's spoken pronunciation with standard pronunciation, and use the Achart Engine to show the comparison chart of the changes of formant to the oral learners so as to make the learners find the problems more intuitively. Besides, the system should give advice on spoken pronunciation based on the relationship between tongue shape, mouth shape and signal formant.

Interface design: The main interface includes four oral training options of vowels, consonants, words and sentences and the learners can choose the items according to their own willingness. At the same time, the resonance peak comparison chart and historical scoring items are added on the main interface to facilitate learners to view. Help options and exit keys are also set. Training score interface elements include pronunciation demonstration, pronunciation following, pronunciation contrast, pronunciation evaluation, main menu, oral demonstration (animation, audio and video, pictures and other forms) and the corresponding text description.

The development of the system is mainly done in the Eclipse integration environment. Specific development and operating environment: PC operating system Windows7 (32bt); Development components: Java JDK 8.0, Eclipse [15] 4.5 (Mars), Android SDK 4.0; Hardware Environment: Glory Play 6X (RAM: 3GB, ROM: 32G, Android 6.0); Programming Language: Java. Figure 2 shows the interface effect.

Figure 2: Oral training system main interface and rating interface.

4 System test results

This study invites three experienced English teachers as score judges and 10 college students as the subjects of the oral English training system. The scoring is based on tongue type, mouth type, pronunciation completeness and clarity, 25 points for each item. The average of the scores given by the three teachers is taken as the final score.

4.1 Speech input test

First, the subjects' speech was recorded and the recognition rate of the speech input system was tested. According to the system instructions, the subjects read after the system of 20 monophthong words, 15 diphthong words and 15 polysyllabic words. The three teachers judged whether the speech input was successful and the results are shown in Table 1.

Word type	Monophthong	Diphthong	Polysyllabic
Total number	20	15	15
Accuracy	100%	100%	96%

Table 1: Speech input success rate of the system.

4.2 Scoring accuracy test

Based on the speech input, scores were given by the system and the three teachers respectively. Suppose the system score of the i-th word was x_i, and the score by the teachers was y_i, the similarity of the two scores was calculated according to the formula $\mu_i = 1 - \frac{|x_i - y_i|}{y_i}$, then, the scoring accuracy of n samples can be calculated based on $\overline{\mu} = \frac{(\mu_1 + \mu_2 + ... + \mu_n)}{n}$, as shown in Table 2.

Word type	Monophthong	Diphthong	Polysyllabic
Total number	20	15	15
Accuracy	97.15%	94.96%	93.62%

Table 2: System test accuracy results.

As shown in Table 2, the accuracy on monophthong word pronunciation reached 97.15%, and that on diphthong words and polysyllabic words reached 94.96% and 93.62% respectively. With the increase of vowels in the words, the pronunciation became complicated, which affected the scoring mode. But, the scoring accuracy reached above 90% on average.

5 Conclusion

With the rapid development of mobile network and the upgrading of mobile network equipment, the concept of mobile learning has been gradually integrated into our life. This paper focused on mobile learning and designed an oral English training system that could be used on Android smartphones. Firstly, we designed the speech signal preprocessing, feature extraction and signal pattern matching of system speech recognition. According to the characteristics of Android system, the dynamic time regulation algorithm with small amount of computation was introduced as the pattern matching algorithm of speech signals. Then, according to the pronunciation characteristics of spoken English, we selected the pronunciation resonance peak as the reference of system scoring, and determined the single reference template as the scoring method. Afterwards, the system functions were designed from the three aspects of pronunciation demonstration, pronunciation imitation and pronunciation evaluation. Finally, the system was tested, the results of which showed that the system had a high success rate in the recognition of the spoken word pronunciation and a high accuracy in spoken English scoring. In general, the system we designed initially met the needs of speech accuracy and scoring accuracy of mobile spoken English training, and provided some points that need attention in pronunciation, which is helpful for spoken English training.

6 References

[1] An L L, Wu Y N, Liu Z, Liu RS (2012). An Application of Mispronunciation Detecting Network for Computer Assisted Language Learning System. Journal of Electronics & Information Technology, 34(9), pp. 2085-2090.

[2] Alamer R A, Al-Otaibi H M, Al-Khalifa H S (2015). L3MS: A Lightweight Language Learning Management System Using Mobile Web Technologies, 2015 IEEE 15th International Conference on Advanced Learning Technologies (ICALT), IEEE, Hualien, Taiwan, pp. 326-327.

[3] Milutinovic M, Bojovic Z, Labus A, Bogdanovic B, Despotovic-Zrakic M (2016). Ontology-based generated learning objects for mobile language learning. Computer Science & Information Systems, pp. 4-4.

[4] Troussas C, Virvou M, Alepis E (2014). Multifactorial user models for personalized mobile-assisted language learning. Frontiers in Artificial Intelligence & Applications, 262, pp. 275-282.

[5] Sharples M, Arnedillosánchez I, Milrad M, Vavoula G (2014). Mobile Learning. R Keith Sawyer, pp. 501-521.

[6] Hu Y, Azim T, Neamtiu I (2015). Versatile yet lightweight record-and-replay for Android. ACM Sigplan International Conference on Object-Oriented Programming, Systems, Languages, and Applications, ACM, Newyork, USA, pp. 349-366.

[7] Deepa D, Shanmugam A (2011). Enhancement of noisy speech signal based on variance and modified gain function with PDE preprocessing technique for digital hearing aid. Journal of Scientific & Industrial Research, 70(5), pp. 332-337.

[8] Takagi T, Seiyama N, Miyasaka E (2015). A method for pitch extraction of speech signals using autocorrelation functions through multiple window lengths. Electronics & Communications in Japan, 83(2), pp. 67-79.

[9] Sahoo T R, Patra S (2014). Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification. International Journal of Image Graphics & Signal Processing, 6(6), pp. 27-35.

[10] Valentini-Botinhao C, Yamagishi J, King S (2012). Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise. Proc. Interspeech., 631-634.

[11] Rathore P S, Boyat A, Joshi B K (2013). Speech signal analysis using Fourier-Bessel Expansion and Hilbert Transform Separation Algorithm. IEEE International Conference on Signal Processing, Computing and Control, IEEE, Solan, India, pp. 1-4.

[12] Dhingra S, Nijhawan G, Pandit P (2013). Isolated speech recognition using MFCC and DTW. International Journal of Advanced Research in Electrical Electronics & Instrumentation Engineering, 2(8), pp. 4085-4092.

[13] Lang F Y, Li X G (2012). Multi-Sensors Information Fusion Based on Momentis Method and Euclid Distance. Advanced Materials Research, 383-390(383-390), pp. 5447-5452.

[14] Yusnita M A, Paulraj M P, Yaacob S, Bakar SA, Saidatul A (2011). Malaysian English accents identification using LPC and formant analysis. IEEE International Conference on Control System, Computing and Engineering, IEEE, Penang, Malaysia, pp. 472-476.

[15] Wang L, Groves P, Ziebart M (2013). Urban Positioning on a Smartphone: Real-time Shadow Matching Using GNSS and 3D City Models. Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation, Nashville Convention Center, pp. 1606-1619.

[16] Thakral S, Goswami D, Sharma R, Prasanna CK, Joshi AM (2016). Design and implementation of a high speed digital FIR filter using unfolding. IEEE, Power India International Conference, pp. 1-4.

[17] Han Z, Wang J (2016). Dynamic feature extraction for speech signal based on MUSIC. Control and Decision Conference, pp. 3770-3773.